without the nicotine patch. *JAMA.* 1994;271: 589–594.

14. Group Health Association of America, Inc. *HMO Industry Profile: 1993 Edition.* Washington, DC: Group Health Association of America, Inc; 1993.
15. Silagy C, Mant D, Fowler G, Lodge M. Meta-analysis on efficacy of nicotine replacement therapies in smoking cessation. *Lancet.* 1994;343:139–142.
16. Fiore MC, Smith SS, Jorenby DE, Baker TB. The effectiveness of the nicotine patch for smoking cessation: a meta-analysis. *JAMA.* 1994;271:1940–1947.
17. Gilchrist V, Miller RS, Gillanders WR, et al. Does family practice at residency teaching sites reflect community practice? *J Fam Pract.* 1992;37:555–563.

# Annotation: Accounting for the Effects of Both Group- and Individual-Level Variables in Community-Level Studies

The Journal devoted space in the May 1994 issue to the design and analytic problems of ecological studies.[1–5] Those same problems are surely germane to the community intervention trials reported in the current issue. A question generally neglected in such trials but addressed in our previous issue is the reconciliation of the common disparities between individual- and group-level analyses.

At the heart of the matter are two basic facts. First, some characteristics are unique to individuals, and while they cannot be captured directly at group level, they can yet impinge on the results of group-level analysis through their distributions and interactions with other individuals and the group. Conversely, other characteristics are unique to groups, and while they cannot be captured at individual level, they too can impinge on the results of individual-level analysis that compares groups.

Susser[4] designated independent, dependent, and associated variables that ostensibly occur at both the individual and group levels, but may act differently at each level; he designated integral and contextual variables as unique to the group level. Koopman[2] emphasized the dependent happenings first described by Ronald Ross[6] and the necessity, in other words, to take analytic account of the group effects when the outcome in one individual affects the outcome in other individuals, a circumstance characteristic of both infectious disease and group behavior. The reader is referred to Koopman's paper and the references contained therein for appropriate nonlinear dynamical methods.[2]

Those responsible for the design and analysis of community intervention trials will also benefit from the study of the statistical properties of cluster sampling (see, e.g., Donner et al.[7] and Donner[8]). An attractive feature of the analysis of cluster samples is that it enables one to fit simple individual-level models without cluster indicators, while adjusting for the correlation between individual responses within clusters. The latter adjustment is of note because in trials with a large number of small communities, a positive correlation between responses of community members decreases the effective sample size and could result in a loss of power. Some of the trials reported in this issue circumvent this problem by analyzing only group-level data, as is appropriate in randomized trials with the community as the unit of randomization.

For the remainder of this annotation, we focus on the changes in regression coefficients that can arise when group-level data are used instead of individual-level data in situations in which different exposures (interventions) occur within groups (communities). The assumption of linear relationships is convenient but not obligatory. We do assume the simplest case of statistically independent responses, given exposures and specific group membership. Note that even within this restricted case, responses of pairs of members from the same group will appear correlated when such pairs are considered from group to group; this is because of the common action of group-level effects.

To reduce the discussion to its elements, we consider three variables only: individual-level exposure, $x$; individual outcome, $y$; and a grouping variable, $G$, which serves to identify the distinct group to which an individual belongs. $G$ usually specifies a finite number of discrete groups but need not do so: we allow groups to be defined by levels of a quantitative variable. To abstract from sampling variability, we shall assume large samples within groups and use the convenient notation of mathematical expectation.

Consider a model that specifies a linear relation at the individual level between $x$ and $y$, given membership in group $G$:

$$E(y|x, G) = \alpha + \beta x + f(G), \qquad (1)$$

where the left side of the model denotes the conditional mean of $y$, given $x$ and $G$. Model (1) is additive with respect to the effects of $x$ and $G$, so that $\beta = \beta_w$ is the within-group regression coefficient of interest, constant across all groups. We specify the group effect $f(G)$ as an entirely arbitrary function to allow for various possibilities. Thus, if there are $k$ distinct groups, a conventional form for the group effect would be

$$f(G) = \gamma_1 I[G = 1] + \cdots + \gamma_{k-1} I[G = k - 1],$$

where $I[G = i]$ is a zero-one indicator for membership in group $i$, and $\gamma_i$ is the difference in mean response $y$ between group $i$ and reference group $k$, given fixed exposure $x$. Alternatively, if groups are defined by levels of a quantitative variable (denoted by $G$ as well), then $f(G)$ could assume the linear form $f(G) = \gamma G$ or any other appropriate function of $G$.

The group effect $f(G)$ is straightforward to observe with individual-level data. When we move to group-level variables $X = E(x|G)$ and $Y = E(y|G)$ by taking averages within groups, model (1) implies that

$$E(y|G) = \alpha + \beta E(x|G) + f(G),$$

or simply that

$$Y = \alpha + \beta X + f(G). \qquad (2)$$

In ecological analysis, the group effects $f(G)$ are generally unavailable, entering (2) as perturbations of the linear relation between group variables $X$ and $Y$. Unlike ordinary error terms in regression models, however, $f(G)$ may be correlated with $X$, leading to a different regression equation. Moreover, $f(G)$ may not even be linearly related to $X$. To investigate these circumstances a bit further, we suppose that the relation between $X$ and $Y$ appears approximately linear; thus, for ecological analysis, one will obtain the best linear predictor of $Y$ given $X$ based on the observed pairs $(X, Y)$. The best linear predictor of $Y$ given $X$, $BLP(Y|X)$, is that linear function of $X$ that minimizes the mean squared error $E\{Y - L(X)\}^2$ among all linear functions $L(X)$, where the expectation is taken with respect to the distribution of $(X, Y)$ across groups.* We distinguish here between $BLP(Y|X)$ and the true regression of $Y$ on $X$, $E(Y|X)$ because the latter may not be a linear function of $X$.** The coefficients of $BLP(Y|X) = a + bX$ are given by $a = EY - bEX$ and $b = \mathrm{cov}(X, Y)/\mathrm{var}(X)$, which agree with the familiar formulas for regression theory in which $E(Y|X)$ is assumed to be linear.

Given model (2), it is easy to show that the best linear predictor of $Y$ given $X$ is $BLP(Y|X) = (\alpha + c) + (\beta + d)X$, where $c$ and $d$ are the coefficients of the best linear predictor of $f(G)$ given $X$, $BLP(f(G)|X) = c + dX$, with $c = Ef(G) - dEX$ and $d = \mathrm{cov}(f(G), X)/\mathrm{var}(X)$. The ecologic coefficient of $X$ is thus $\beta_e = \beta + d$, and we see that $\beta_e = \beta_w$ if and only if $d = 0$. This occurs when either $f(G)$ is identically zero, or the group effect $f(G)$ is uncorrelated with $X$. The first condition occurs when there are no group effects on individual outcome $y$ in model (1) given individual exposure $x$—that is, when outcome is conditionally independent of group membership given exposure, even if $G$ is highly correlated with $X$. For example, if $E(y|x) = \alpha + \beta x$ in model (1) and $G$ is defined by a grouping of individuals in subintervals of $x$, the ecological coefficient of $X$ agrees with $\beta$. The second condition would occur, for example, if groups were constructed as random samples and thus were uncorrelated with $X$. Of course, one recognizes these two conditions as being sufficient to ensure that $G$ is not confounding between $x$ and $y$ in the linear model (1). In cases where neither condition obtains, coefficients $\beta_e$ and $\beta_w$ generally differ.

If we elaborate on model (1) to allow $\beta$ to depend on group, say $\beta = \beta(G)$, then the shift in the ecological coefficient of $X$ contains an additional component. Let $\delta(G)$ denote the effect modification of group $G$ on exposure: $\delta(G) = \beta(G) - \beta_w$, where $\beta_w$ is the weighted average of within-group slopes.* Then the ecological coefficient is $\beta_e = \beta_w + d$, where the shift is now $d = \mathrm{cov}(\{\delta(G)X\} + f(G), X)/\mathrm{var}(X)$. The additional component $\mathrm{cov}(\delta(G)X, X)/\mathrm{var}(X)$ will not generally equal zero unless the group-effect modifications $\delta(G)$ are uncorrelated with both $X$ and $X^2$. □

***Bruce Levin***
***Consulting Editor for Statistics***

*Here and below, expectations, variances, and covariances of group-level variables across groups are written $EX$, $\mathrm{var}(X)$, $\mathrm{cov}(X, Y)$, etc. In calculating these quantities, groups are often weighted proportional to their size. For example, with $k$ discrete groups of size $n_i$ $(i = 1, \ldots, k)$, $EX = \Sigma_i n_i X_i / \Sigma_i n_i$. For continuous groups, the moments are weighted by the probability density function of $G$.

**What characterizes the best linear predictor is that the residuals $Y - BLP(Y|X)$ have mean zero and are uncorrelated with $X$. These are weaker conditions than the usual requirement in linear regression that the error term have zero conditional expectation given $X$.

*The weights are proportional to $\mathrm{var}(x|G)$: $\beta_w = E\{\beta(G)\mathrm{var}(x|G)\}/E\{\mathrm{var}(x|G)\}$. A certain linear combination of $\beta_w$ and $\beta_e$ produces the slope, $\beta_t$, of the best linear predictor of $y$ given $x$ ignoring $G$. The relation is $\beta_t = V\beta_w + (1 - V)\beta_e$, where $V = E\{\mathrm{var}(x|G)\}/[E\{\mathrm{var}(x|G)\} + \mathrm{var}(X)]$ gives the proportion of total variance in $x$ accounted for by the average within-group variance.

## References

1. Poole C. Ecologic analysis as outlook and method. *Am J Public Health.* 1994;84:715–716.
2. Koopman JS, Longini IM Jr. The ecological effects of individual exposures and nonlinear disease dynamics in populations. *Am J Public Health.* 1994;84:836–842.
3. Schwartz S. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *Am J Public Health.* 1994;84:819–824.
4. Susser M. The logic in ecological: I. the logic of analysis. *Am J Public Health.* 1994;84:825–829.
5. Susser M. The logic in ecological: II. the logic of design. *Am J Public Health.* 1994;84:830–835.
6. Ross R. *The Prevention of Malaria.* 2nd ed, rev and enl. London, England: John Murray; 1921.
7. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *Am J Epidemiol.* 1981;114:906–914.
8. Donner A. Sample size requirements for stratified cluster randomization designs. *Stat Med.* 1992;11:743–750.

# Annotation: Confounding in Epidemiologic Research

In this issue, Simoes et al. used data from the 1990 Behavioral Risk Factor Surveillance System to examine the association between leisure-time physical activity and dietary fat intake in a population-based sample of US adults.[1] The authors observed that physical activity and dietary fat were inversely correlated, independent of other sociodemographic and health characteristics. While they could not assess fat intake as a percentage of total caloric intake because of the limited nature of their dietary information, the inverse relation with percentage fat intake would likely have been more pronounced since physically active individuals consume more calories than sedentary ones. Based on these data, the authors recommend that epidemiologic studies of either factor (physical activity or diet) consider controlling for the other, since the two may potentially confound each other.

The relevance of this finding relates to the importance of confounding when interpreting the results from an epidemiologic study. In any study, what we assess is whether an exposure of interest is associated with a particular outcome. The presence of a statistical association, however, in no way implies that the observed relation is one of cause and effect; yet, from a public health standpoint, the judgment of a cause-effect relationship is the primary objective in epidemiology. To decide so is neither simple nor straightforward: it requires not only an assessment of the validity of the results seen in an individual study but also a judgment based on the totality of evidence.

In order to assess whether the findings of a study represent a valid (or true) association, we have to determine the likelihood that alternative explanations—chance, bias or confounding—could account for the results.[2] Confounding refers to a mixture of effects between the exposure and outcome studied and a third factor (the confounder) that is associated with the exposure and, at the same time, an independent risk factor for the outcome. If not controlled for, confounding can lead to either the observation of an artifactual association between exposure and outcome, or, conversely, the observation of no association when one truly exists. For example, the evidence from epidemiologic studies, many of which did not control for fat intake, suggests that physical activity is inversely related to risk of coronary heart disease.[3] Dietary fat also has been suggested to be an inde-

**Editor's Note.** See related article by Simoes et al. (p 240) in this issue.